Cognitive Process Modeling of Spatial Ability:
A Construct Validity Study of an Assembling Object Task

Jennifer L. Ivie

Submitted to the Department of Psychology
and the Faculty of the Graduate School of the University of Kansas
in partial fulfillment of the requirements for the degree of Master of Arts

Redacted Signature

Chair

Redacted Signature

Redacted Signature

Date submitted: 12/17/2002

ABSTRACT
Jennifer L. Ivie, M.A.
Department of Psychology
University of Kansas

The purpose of this study was to examine the cognitive processes involved in completing a spatial task in which a participant must mentally assemble a two-dimensional objects. These tasks are used to measure spatial ability on tests such as the Revised Minnesota Paper Form Board Test. Two studies were completed to support a cognitive processing model, previously proposed by Embretson and Gorin (2001), for stages a participant must go through to solve this problem type. In the first study, data from a large group of students from the University of Kansas was used to discover what variables could be manipulated within each item to effect item difficulty and mean response time. Multiple regression models and linear logistic latent trait models were used to measure the impact of each variable on its respective cognitive processing stage. Finally, an eye tracker study was done on ten students from the University of Kansas to further support the proposed cognitive processing model. A qualitative analysis of the data generally supported the proposed cognitive model, but also indicated necessary revisions.

Acknowledgments

The author would like to thank Dr. Susan Embretson, the advisor for this thesis, for her guidance and assistance in completing this project. The author would also like to thank the other committee members, Dr. Doug Denney and Dr. Janet Marquis, for their willingness to join at such late notice and for their support throughout the final stages of this process.

The author would like to thank Dr. Joanna Gorin for the previous work done on this project as well as for her undying support and assistance while finishing her degree. In addition, the author would like to thank Lesa Hoffman and Dr. James Bovaird, IV, for their support and statistical assistance.

Finally, the author would like to thank her family and friends for believing in her and for continually pushing her throughout the many years leading up to this point.

Cognitive Process Modeling of Spatial Ability:

A Construct Validity Study of an Assembling Objects Task

Table of Contents

List of Tables

List of Figures

List of Appendices

Cognitive Process Modeling of Spatial Ability:
A Construct Validity Study of an Assembling Objects Task

Cognitive psychology research has placed much focus on developing tests of cognitive abilities and intelligence that are widely used in educational and occupational settings. This research helps develop more efficient test design methods as well as helping develop more valid tests. In line with this type of research, this study examines the cognitive processes involved in spatial processing tasks, in particular the Assembling Objects (AO) task found on the Revised Minnesota Paper Form Board Test (RMPFBT).

A summary of previous research in cognitive abilities testing and spatial processing is included to provide a background for this study. Multiple models for the cognitive processing of different spatial tasks are presented. For this study, one of these models is proposed for solving these particular AO items. Also, a set of variables hypothesized to affect item difficulty and response time are described. Finally, the introduction concludes with an explanation of the design of this study.

History of Intelligence Testing

Standardized testing has been a focus of intelligence research for many decades. Psychometricians have developed many measures of "intelligence," "general cognitive ability," "scholastic ability," etc. Testing has played a central role in learning the nature of intelligence and other cognitive abilities. This use of tests measuring intelligence or cognitive functioning has proven useful in many educational and occupational settings (Sternberg, 1994).

In 1905, Alfred Binet and Theophile Simon developed the first practical intelligence test. This test consisted of a variety of tasks of cognitive ability and dexterity, ranging in level of complexity. The revision in 1908 provided each individual with a summary score that allowed for the first measure of the intelligence quotient (IQ). Most tests consisting of a battery of tasks developed since have also provided an overall score considered a measure of intelligence (Sternberg, 1994).

Yet, other tests have been developed using a single task to measure cognitive ability or a single facet of intelligence rather than a battery of multiple cognitive tasks. Some of these tests have been considered to measure intelligence wholly or just one specific factor of intelligence. One single-task test that is highly correlated with general intelligence is the Progressive Matrices test first developed by J. K. C. Raven (1962) and later revised by S. E. Embretson (1984). This test resulted from a study examining another test of spatial analogies developed by C. Spearman in 1927, suggesting that optimally a test of general intelligence should involve "relations and correlates" of parts of a whole (Sternberg, 1994). There have been many other tests developed since to measure different aspects of spatial reasoning.

Background of Spatial Ability

The most accepted definition of spatial ability is that it is the ability to "generate, retain, retrieve, and transform well-structured visual images (Lohman, 1988)." While many tasks measure spatial ability as a single ability, it is believed that spatial processing is a combination of many abilities.

There are many theories as to the dimensionality of intelligence. Galton (as cited in Sternberg, 1994) was the first to propose one general all-encompassing intelligence factor. Spearman (as cited in Sternberg, 1994) followed with his two-factor theory of intelligence, where intelligence is divided into a general factor (*g*) and specific factors (*s*) related to different tasks or abilities. Under this theory, spatial aptitude would be classified as a specific factor of intelligence. Overall researchers believe most evidence supports a theory of the multidimensionality of intelligence.

In the 1970s and 1980s, Gardner (as cited in Sternberg, 1994) developed his Multiple Intelligences theory claiming that everyone has the potential to cognitively develop within a set of seven intellectual faculties: linguistic, logical-mathematical, musical, spatial, bodily-kinesthetic, interpersonal, and intrapersonal. Spatial intelligence is reflected in the ability to create mental representations over local forms of space or over a more large scale space.

Thurstone (1931) proposed a theory of primary mental abilities. This theory too divides intelligence into seven primary abilities: verbal comprehension, word fluency, number facility, spatial visualization, inductive reasoning, memory, and perceptual speed. He defined spatial ability as the "facility in spatial and visual imagery" and perceptual speed as the "facility in finding or recognizing particular items in a perceptual field." Thurstone also found the spatial factor to be fairly associated with verbal comprehension and inductive reasoning.

Results from studies of tests measuring intelligence as defined by the afore-mentioned theories tend to support the three-stratum theory of cognitive abilities

again supporting the multidimensionality of intelligence. At the first and lowest stratum are the primary abilities or multiple intelligences. At the second or middle stratum are several second-order factors (e.g., fluid and crystallized intelligence). And, at the top or third stratum is Spearman's *g*, or general intelligence. Thus, the more general factors are found at the top, whereas the more specific factors are found at lower levels (Sternberg, 1994).

Another multidimensionality theory of human intelligence that has been proposed describes intelligence as having a radex form. Lohman (2000) suggested a hypothetical radex map of cognitive abilities. To picture a radex map, one could think of a slice of a tree trunk. This type of map consists of a circle divided into smaller circles with the same center. Emerging from the center are radii that divide the circle into pie pieces. In a radex map, the closer to the center of the circle a task falls, the more complex the task is. Also, the radii usually divide the map into sections of related tasks. In Lohman's hypothetical radex, there are three sections—mathematical reasoning, verbal reasoning and spatial reasoning. At the center of his radex is the Raven's Progressive Matrices task. The Assembling Objects task and Paper Form Board task both fall closer to the outside of the circle of the radex, suggesting that these tasks are often easier than other tests of spatial ability.

Studies by Lohman (1979) demonstrated three distinct spatial factors. The first of these factors is Spatial Orientation. This involves the ability to imagine how the stimulus would appear from another perspective. Measures of this ability require the movement of three dimensional items in space and are often used in engineering

fields. The second factor is Spatial Relations, or the ability to engage rapidly and accurately in necessary mental rotation for comparing the identity of a part of the stimuli. Tests like Thurstone's Primary Mental Abilities Space Test developed in 1931 load on this ability (as cited in Sternberg, 1994). The final factor, Spatial Visualization, is the ability to manipulate internal parts of the stimuli configuration or the ability to fold or unfold stimulus. Lansman, Donaldson, Hunt and Yantis (1982) found Spatial Visualization to be highly correlated ($r = .78$) with ability to perform mental rotations. Likert and Quasha's (1970) Minnesota Paper Form Board Test, which requires a participant to mentally reconstruct images that have been separated into pieces, is an example of a test that loads on Spatial Visualization. Measures of Spatial Relations are usually speeded tests, whereas measures of Spatial Visualization usually measure both speed and accuracy (Pellegrino, Mumaw & Shute, 1985).

Along these same hierarchical concepts, Carroll (1993) identified five major factors involved in spatial reasoning. The first factor, Visualization, is the ability to manipulate visual patterns without regard to the speed of task solution. The second factor, Speeded Rotation, is the speed with which an individual can manipulate relatively simple visual patterns by mental rotation, transformation, etc. The third factor, Closure Speed, is the speed with which an individual can apprehend and identify an unknown visual pattern that is disguised or obscured in some way. The fourth factor is Closure Flexibility—the speed with which an individual can find, apprehend, and identify a known visual pattern that is disguised or obscured in some way. Finally, the fifth factor, Perceptual Speed, is the speed with which an individual

can find a known visual pattern, or accurately compare one or more patterns, in a visual field where the patterns are not disguised or obscured in any way. Both Speeded Rotation and Perceptual Speed are factors that seem to be measured by an Assembling Objects task like the Revised Minnesota Paper Form Board Test.

While there are many measures used in research on spatial ability, very few assess all factors mentioned above. Most tests only measure one or two of these hypothesized factors. Research will continue to examine the impact of these factors on the mental processes involved in solving spatial tasks. Another question researchers are beginning to ask along with what factors are involved in spatial processing is that of the strategies used in solving spatial tasks. How do individuals differ in the processes used to complete these tasks?

Individual Differences in Spatial Cognition

There have been four proposed hypotheses that have directed research in explaining individual differences in spatial processing ability. The first hypothesis is that individuals differ in how fast they can perform analog transformations. A second hypothesis is that there are individual differences in the level of skill required for generating and retaining mental representations that preserve configuration information. A third hypothesis is that individuals differ in how much visual-spatial information they can maintain in an active cognitive state. The fourth hypothesis is that there are differences in the sophistication and flexibility of strategies individuals use in solving these tasks (Lohman, 2000).

In line with the fourth hypothesis, Just and Carpenter (1985) proposed that individuals use a cognitive coordinate system for processing spatial relationships. Their research supports the premise that mental rotation occurs around cognitively determined axes, much like the Cartesian coordinate system used in mathematics. These axes cross at a center or origin occurring at the most detailed part of the object to be rotated, just as the *x*- and *y*-axes cross at a zero point. Their studies also demonstrated that the use of a cognitive coordinate system affects recognition and information retrieval, as well as spatial transformations, such as mental rotation.

Four main strategies using a cognitive coordinate system for solving spatial transformations have been illustrated. The first and most frequently used strategy is mental rotation around standard mathematical axes (i.e., the *x*-, *y*-, and *z*-axis). The second strategy is to mentally rotate around arbitrary axes defined by the individual in terms of the task at hand. The third strategy is to rotate around object-defined axes invariant with the object's orientation in space. The final strategy is to code the observer's position with the object's position within the cognitive coordinate system (Just & Carpenter, 1985).

Going along with the second hypothesis, stating that differences lie in skill level related to configuration preservation, research in cognitive psychology has led to two theories of mental representations of spatial knowledge. One theory proposed that spatial knowledge could be represented in a literal manner (Kosslyn, 1980), or a manner in which the original structure or configuration is preserved (Anderson, 1983). A second, more abstract theory is that the meaning or interpretation of the

original structure or configuration is preserved in the mental representation (Anderson, 1983; Kosslyn, 1980).

Until recently, intelligence measures were only able to give an overall score on a particular ability and were unable to discover the underlying differences between individuals' strategies and ability levels. Researchers are beginning to design ways to measure these differences, either through self report or open-ended questions requiring participants to work out problems and not just choose an answer from multiple distractors.

Measures of Spatial Ability

Measures of spatial aptitude have become important for predicting scholastic and occupational success in technical fields where verbal and quantitative skills are not as effective (Lohman, 1979). Some of these fields include: architects, engineers, draftspersons, cabinetmakers, mechanics, airplane pilots, air traffic controllers, etc. (McGee, 1979). Eighty-four different job categories that require high spatial abilities are listed by the U.S. Employment Services (1957). High scores on spatial ability measurements have also been linked to creativity in the arts, sciences and mathematics (West, 1991).

Humphreys, Lubinski and Yao (1993) did a longitudinal study of high school students where they took a measure of their Spatial-Math ability and their Verbal-Math ability and then tracked their career paths for 11 years following graduation. They found evidence to support the theory that a measure of spatial visualization should be added to the already used standardized testing for admission into higher

education institutions. They concluded that many applied fields such as scientific research, engineering and art are possibly losing many capable individuals due to the current state of assessment for placement.

Four types of measures of spatial abilities have been used: performance tests, paper-and-pencil tests, oral tests, and film or dynamic computer-based tests. In 1916, the first Binet intelligence test included three tests of spatial ability. These three tests consisted of form board, block manipulation and paper-folding tasks (Sternberg, 1994). Binet and Simon (1916) were the first to include form board, block manipulation, and paper-folding tasks on their performance test. Many of these tasks are still used today in measuring children's performance or nonverbal intelligence.

One example of a spatial processing test, the Primary Mental Abilities (PMA) Space Test, was developed by Thurstone in 1931. On this test the participant is required to study an object, then search through the five choices and find the one that is the same shape or form but rotated 0 to 300 degrees. The stem object is a letter from the alphabet or some unfamiliar line drawing. The PMA continues to serve as a prototype for spatial processing tasks that require mental rotation and spatial relations (Mumaw, Pellegrino, Kail & Carter, 1984). In 1934, in the first administration of this test, 240 students were given a fifteen hour test consisting of fifty-six batteries. Thirteen of the test variables had at least a .40 positive factor loading on one factor assumed to be a spatial factor. All of thirteen of these variables were related to visual or spatial processing (Sternberg, 1994).

Another test of spatial ability, the Minnesota Paper Form Board Test designed by Likert and Quasha (1970) is an example of a measure of Spatial Visualization. In this test, a participant is required to assemble pieces in the stem to match one of the alternatives (see Figure 5, page 27). Pellegrino, Mumaw and Shute (1985), examined this test to see what dimensions underlie task difficulty and errors. Rather than using the original items from the test, they used a variant of the task. They placed the stem next to an assembled version of the pieces (see Figure 1) and the participant was supposed to decide whether the pieces match the assembled object. The pieces could vary from the assembled object in five ways: rotated and displaced, rotated, displaced, separated or holistic (or not separated).

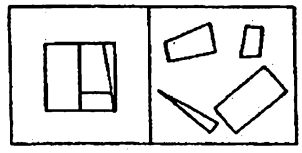

Figure 1. Sample item used in the Pellegrino, Mumaw and Shute (1985) study.

Pellegrino and colleagues proposed the cognitive processing model seen in Figure 2 for how subjects solve these items. In this model, a participant must first encode some piece from the stem. Then, the participant uses that piece to search for a matching piece in the assembled version. Sometimes he or she must mentally rotate or displace this piece to make a match. If that piece does not match, the participant can conclude that the assembled version is different. But, if the participant concludes

that the piece does have a matching piece in the assembled version, the participant must then choose another piece from the stem and go through the same steps until all pieces have been exhausted. If the participant finds a match for all pieces in the stem, he or she can then conclude that the assembled version is not different from the pieces in the stem.

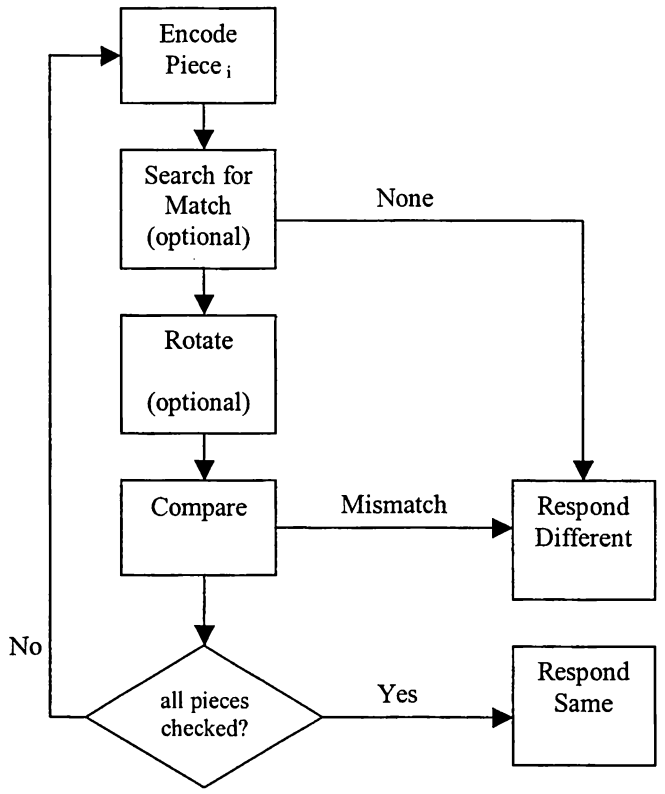

Figure 2. Pellegrino, Mumaw and Shute (1985) cognitive processing model for items like that seen in Figure 1.

Embretson and Gorin (2001) designed a program to create on-the-fly Assembling Object (AO) items, similar to those found on the Revised Minnesota Paper Form Board Test (RMPFBT) (see Figure 3). For these items, there is a stem consisting of a shape cut into two to five pieces. Then the pieces are separated and possibly rotated or displaced. The participants are required to find the key, which is an assembled version of the stem (just as in the task used by Pellegrino, Mumaw and Shute), among four choices.

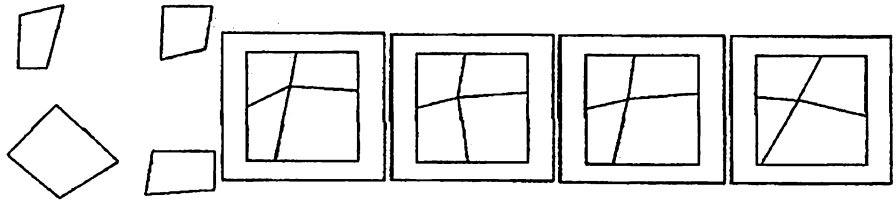

Figure 3. Example of a generated item from Embretson and Gorin (2001).

Embretson and Gorin (2001) proposed a processing model for these item types (see Figure 4, page 15) similar to that proposed previously by Pellegrino, Mumaw and Shute. The model they proposed will be the one used in the current study and will be discussed later.

Background of Cognitive Process Modeling

Many cognitive researchers use information-processing models to analyze the mental processes involved in problem solving of various tasks. These processing models, like that seen in Figure 3, usually consist of one or more sequences of steps

or stages in which cognitive operations are performed. Some of these models are simple constructions that reflect the functioning with only one or two parts. Other models are compound mathematical models combining multiple simple models used to define processing of more complex tasks (Lohman, 2000).

Discovering the processes involved in solving a task is not only useful in learning more about cognition but is also useful in test design. By modeling the cognitive processes involved in completing particular tasks, items can then be designed that are consistent with these processes. This provides for more valid test construction and allows for research on what variables make each stage more or less difficult (Embretson, 1998).

The role of cognitive theory is limited by construct validation. Theoretically, an ability construct must be defined after a test is designed. However, there is an alternative view of construct validity. Construct representation can be used to develop alternative cognitive processing models for a particular ability to be measured. After the test has been designed, a correlation between it and other measures can be found to measure the validity of the test. This method for test development is known as the Cognitive Design System approach (Embretson, 1999) and is the method used in the current study.

Proposed Cognitive Model for Spatial Processing

Cooper and Shephard (1973) gave evidence for a four stage processing model for solving mental rotation tasks on the Primary Mental Abilities (PMA) Space Test. As mentioned earlier, this test consists of a stem object, a key choice and four

distractor choices. The participant is required to find the choice which is the stem object rotated 0 degrees to 300 degrees. In their model, the first stage requires encoding the stimuli. The next two stages are cyclical. First, the participant must rotate the stem object and then compare it to each of the choices. Finally, the last stage is a motor response or execution after confirmation of the key.

A similar model for processing Assembling Object (AO) items was proposed by Embretson and Gorin (2001). In this task there is a stem consisting of a shape cut into pieces, the pieces are then separated, and sometimes rotated or displaced. The participants are required to choose an answer from five choices which is the assembled version of the pieces from the stem. The proposed model is a three stage model. These three stages include Encoding, Falsification, and Confirmation. This model can be seen in Figure 4. In this three stage model, we can see the steps that theoretically a person takes to solve an Assembling Object problem. First, the person must encode the stimulus or stem. This requires processing each individual piece of the stem, its general shape and size, the number of sides, and the total number of pieces in the stem. Next, the person falsifies each alternative using one or more of these encoded pieces from the stem. The third stage, Confirmation, requires falsifying any non-falsifiable distractor(s) and verifying that every piece in the key matches the stem.

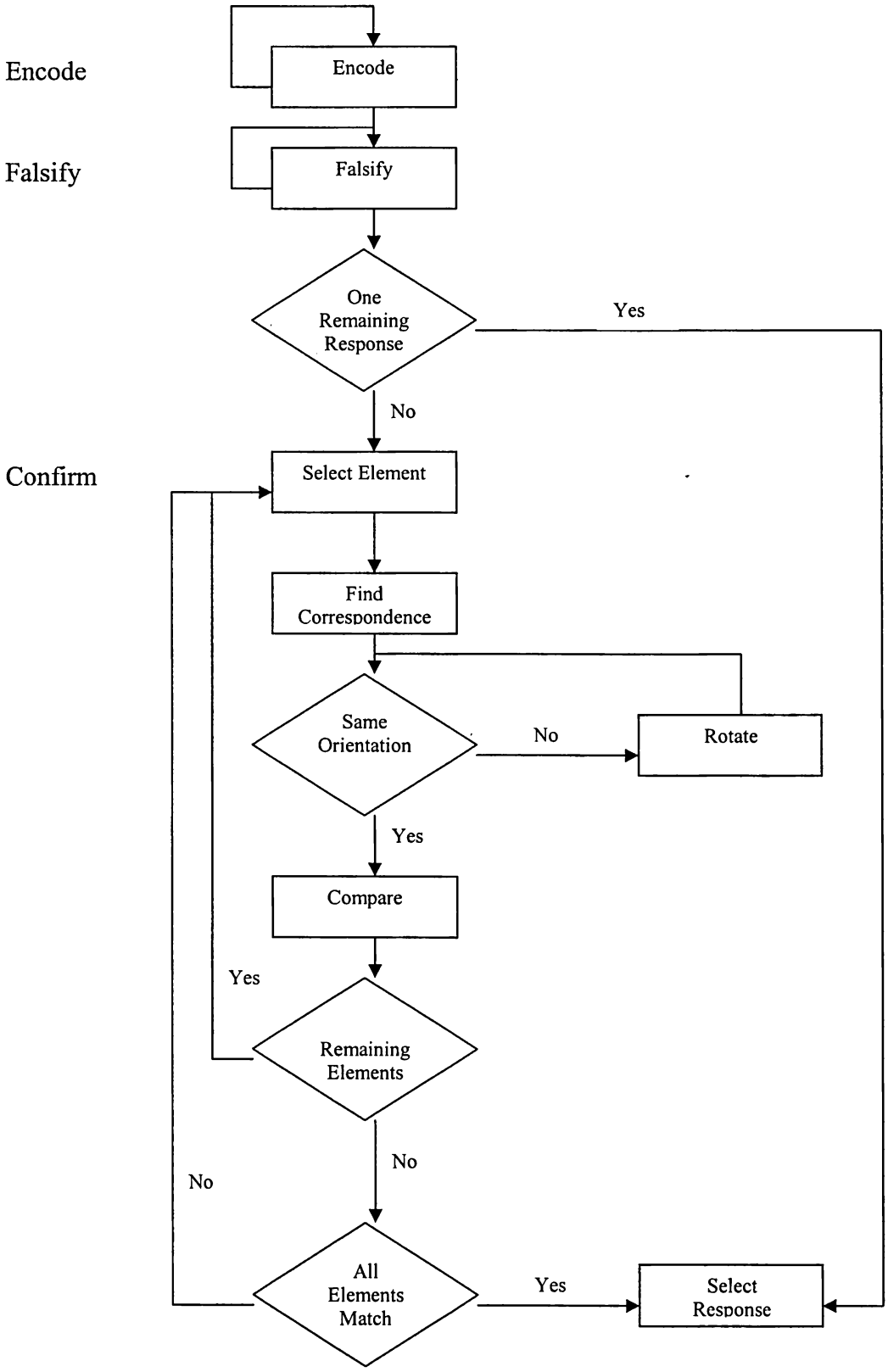

Figure 4. Cognitive processing model for solving Assembling Objects items.

There are multiple ways to evaluate what variables or item attributes influence the stages in the processing model. After defining the attributes, researchers can analyze the contributions of these attributes through multiple regression or linear logistic latent trait modeling. The advantage of the latter is that it takes into account the raw data from the participants as well as the item attributes, whereas the former only takes into account predefined IRT parameters such as item difficulty.

Linear Logistic Latent-Trait Model

In the 1980's, a new measurement system began to emerge to replace classical test theory (CTT). Item response theory (IRT) became a dominant topic of study among measurement specialists. IRT is based on two ideas. First, the performance of an examinee on a test item can be predicted by latent traits of abilities. Secondly, the relationship between an examinee's performance on an item and the traits underlying item performance can be described by a function called an item characteristic curve (ICC). This function shows that as the level of the trait increases, the performance on the item increases as well (Hambleton, Swaminathan & Rogers, 1991).

More than one item response model exists. These models differ mathematically by the number of parameters (e.g., item difficulty, item discrimination, individual's ability level, etc.) specified in the model. Also, these models are falsifiable, that is, a particular model may or may not adequately predict the test data (Hambleton, et al., 1991).

The Linear Logistic Latent Trait Model (LLTM) is one particular IRT model. LLTM was developed in the early 1970's by Fischer to take into account item content

in the prediction of item success. If scorable content factors for each item can be specified, LLTM tests the impact of these factors on item difficulty using the following equation (Embretson & Reise, 2000):

$$P(X_{is} = 1 | \theta_s, \tau_k) = \frac{\exp(\theta_s - \Sigma_k \tau_k q_{ik})}{1 + \exp(\theta_s - \Sigma_k \tau_k q_{ik})}$$

where

q_{ik} = value of content factor k in item I, where q_0 = unit vector

τ_k = weight of content factor k in item difficulty

θ_s = ability level of subject s

LLTM allows researchers to measure the contribution of predefined predictors. These attributes can be predictors of both item difficulty and response time. Item difficulty is usually focused on by most IRT psychometricians, but response time has been a measurement focused on by cognitive psychologists for decades.

Response Time Research

Reaction time is defined as the amount of time it takes a subject to respond to a particular stimulus. It has been studied for decades as a dependent variable that can be affected by manipulating the stimuli and/or the presentation of the stimuli. In testing, because it takes longer to solve a problem than to "react" to stimuli, reaction time is most often referred to as response time (RT), or the time it takes a participant to process and respond to the problem question or item. Analyzing how long it takes

a participant to respond and why it takes a certain amount of time to complete a problem has become more prevalent in testing research in more recent decades.

In 1868, Donders (as cited in Luce, 1986) suggested that the amount of time it takes a person to complete one stage of processing a particular item type could be inferred through experiments where only factors influencing that stage are manipulated. Jastrow (as cited in Luce, 1986) later suggested that if cognitive processing of a particular item type is a process that could be broken down into stages, then each stage would take a particular amount of time to complete. Research in cognitive process modeling is beginning to look into the time it takes to complete each stage in a model, though currently it is still difficult to accurately gauge this sort of measurement.

Galton (as cited in Sternberg, 1994) believed response time to be an important measure of intelligence. He administered RT tests to large groups of people. Other researchers who utilized this method of measuring intelligence and abilities included Spearman, Burt and Cattell (Sternberg, 1994), though they looked at more general intelligence and abilities. Other researchers have begun observing response time as an independent variable in spatial processing tasks.

Shephard and Metzler (1971) developed a test to measure mental rotation as a function of response time. In this task, participants were required to make same-different judgments between two three-dimensional objects. These objects differed in their orientation as well as in whether they were the same shape or mirror images of each other. They found that RT was linearly related to amount of mental rotation

required to make a decision. As the amount of mental rotation required increased, so did the mean RT for solving that item.

Egan (1979) found mental rotation RT to be minimally correlated with spatial ability. He did find, however, that the amount of time a subject takes to execute the encoding stage of processing was correlated with spatial ability.

Response time is a measure that seems to reflect a person's ability level. E. Hunt (1982) summarized several studies as showing a different pattern of RT for individuals of different spatial ability levels. He noticed that high ability participants tended to take more time encoding the stem, whereas lower ability participants moved quickly from the stem to the alternatives and making a choice.

These types of studies help to determine at which ability levels subjects can perform these items at which corresponding speeds. While technology is advancing towards being able to measure the time required to complete each individual stage of the particular cognitive model being studied, one way to qualitatively look at the breakdown of these stages in cognitive processing of AO items is by using an eye tracker.

Using an Eye Tracker

The eyes are the most active sense organ in the human body, continually moving as they scan the details of the visual world. These movements are called saccades, and typically one saccade is made every 250-350 msec when searching for a specified target (Bertera & Rayner, 2000). These saccades are so fast they occupy only about 10 percent of the viewing time, with the other 90 percent taken up by eye

fixations. A fixation is when the eye is aimed at a fixed point in the visual field (Noton & Stark, 1971). In other words, a fixation is what occurs between every saccade. These fixations usually last about 250 msec (Hoffman, 1998).

The theory supported by most literature on how the human mind creates a mental representation of a visual stimulus is that the representation is created through a step-by-step process, looking at individual parts of the stimulus and internalizing the pieces to represent the whole stimulus. Noton and Stark (1971) found that the brain focuses cognitively on the angles and principle features of the visual stimulus. They also found that when a participant recognizes the original stimulus through matching, he or she uses the same fixation pattern when looking at the matched item as he or she did when creating an internal representation of the original stimulus. Thus, the participant fixates on the same corners, or angles, or lines of the matched item, in the same order as he or she did while encoding the original item.

Yarbus (1967) demonstrated that the patterns of eye fixations produced are influenced by the goals and interests of the viewer as well as the properties of the task being viewed. Because attention shifts can occur much more rapidly than changes in fixation, Hoffman and Subramaniam (1995) found that spatial attention can be used to select the location for the next fixation. Spatial attention is the act of allocating mental resources to something in space (Palmer, 1999). The oculomotor readiness hypothesis states that the same neural circuitry (e.g., superior colliculus) mediates both attention and saccades. This hypothesis makes two predictions: 1) attentional enhancement to a future fixation location should be produced by preparing to make a

saccade and 2) attending to a location should result in a quick saccade to that location (Hoffman & Subramaniam, 1995). There is evidence that these saccades are programmed in the brain in a hierarchical manner, the first being direction, followed by an amplitude parameter (Hoffman, 1998).

Eye tracking has become a popular method for studying visual and cognitive processes. Some examples of domains in which eye tracking has been used include image scanning, driving, arithmetic, analogy, and reading. Salvucci and Goldberg (2000) proposed a taxonomy of algorithms for performing fixation identification. This taxonomy consists of five basic types of algorithms that fall under either spatial or temporal categories. The three types of spatial algorithms are velocity-based, dispersion-based, and area-based. The two types of temporal algorithms are duration sensitive and locally adaptive.

Velocity-based algorithms are based on the fact that fixation points have low velocities and saccades have high velocities. Any point-to-point velocity that falls under a specified threshold is defined as a fixation, while those that exceed the threshold are defined as saccades. Velocity is measured as a distance-time ratio by most eye tracking machines. Dispersion-based algorithms assume that fixation points generally occur near each other. With this type of algorithm, all fixations that fall within a defined dispersion area are compressed into one fixation point. Area-based algorithms identify points within given areas of interest representing relevant visual targets. A relevant visual target is sometimes termed a look zone. Duration sensitive algorithms are based on the fact that fixations are rarely less than 100 msec, usually

ranging between 200 and 400 msec. Local adaptive algorithms allow for temporally adjacent points to be interpreted as a single data point (Salvucci & Goldberg, 2000).

The eye tracking machine used for the current study utilizes an area-based algorithm. It measures the amount of time spent fixated on a point within a pre-defined look zone. The look zones were defined by the pieces in the stem and the assembled choices.

Focus of Current Study

The current study will attempt to generalize the Embretson and Gorin (2001) processing model (see Figure 4, page 15) to a different collection of AO items—namely the Revised Minnesota Paper Form Board items. Data for calculating item difficulty can easily be gathered using a paper-and-pencil version of a test, considering the substantial amount of subjects that can participate at any given time. Response time data collection is more limiting, because response time data is best collected using a computer program that records the beginning and ending times for each item. This limits the amount of data collected due to the number of computers available to administer a test at any given time.

There are two methods for testing the proposed processing model. After choosing the variables that affect each stage of the processing model, regression models will be used to analyze the influence of each variable on the processing model. Another way to test the model is to use the IRT method, linear logistic latent trait modeling. This modeling technique uses the variable scores as well as the raw

data to estimate the processing model parameters giving a more accurate estimate of the model parameters, with smaller standard errors.

Finally, an eye tracker study can help to support the results found through modeling. Data can be studied qualitatively to see if the fixation patterns support the order of the stages of the processing model. For example, in the current study, a fixation pattern that would support the model could begin with the eye fixating on different pieces of the stem and then picking a piece and looking back-and-forth between that piece and the alternatives and finally focusing on one particular alternative and choosing it for the answer.

Study 1

The goals of Study 1 include defining variables that could possibly affect item difficulty for these AO items and modeling item difficulty and response time to generalize the proposed processing model to items of this type. The proposed cognitive model being tested is the three-stage model proposed by Embretson and Gorin (2001). This model consists of the Encoding, Falsification, and Confirmation stages (see Figure 4, page 15).

Method

Design

To operationalize the cognitive model, items were scored on the following variables based on those outlined by Embretson and Gorin (2001): the number of pieces in stem, the total number of edges on all pieces in the stem, the maximum number of edges in one piece in the stem, the number of curved pieces in the stem,

the number of pieces with verbal labels in the stem, the number of pieces falsifiable by gross size, shape and angular disparity per alternative, the number of non-falsifiable distractors (NFD) (that is, alternatives that cannot be falsified at first glance), the number of expected cycles necessary to falsify NFDs, the number of pieces mismatched by size between the stem and the NFD(s), the number of pieces mismatched by small angular disparity between the stem and the NFD(s), the number of pieces that must be rotated to match the stem to the key, the number of displaced pieces between their position in the stem and their position in the key, and key position (i.e., the distance of the key from the stem).

The calculation of number of expected cycles necessary to falsify a NFD must further be explained. This is the sum of the probability of the cycle occurring times the success of the cycle times the number of the cycle. The probability of the first cycle occurring is calculated as the ratio of mismatched pieces between the stem and NFD to total number of pieces in the stem. The probability of each subsequent cycle occurring is calculated as a sum of the preceding products of probabilities and successes. For example, if there are 2 mismatching pieces and 4 pieces in the stem, the expected number of cycles would be calculated as follows:

$$E(N_{cycle}) = 1(1)(2/4) + 2(.5)(2/3) + 3(1-(2/4 + (.5)2/3))(1).$$

A new variable not scored by Embretson and Gorin (2001) was the expected number of distractors falsified by any given piece in the stem. Given that each piece in the stem has an equal probability of being chosen, the expected number of

distractors falsified was calculated by the following equation:

$$E(N_{falsified}) = \Sigma_p(1/N_{pieces})(N_{p,false}).$$

In other words, the expected number of distractors falsified is equal to the sum of the probability of selecting a piece *p* times the number of distractors falsified by piece *p*. A list of the variables that fall under each processing stage can be seen in Table 1.

Table 1

List of variables scored per item on the RMPFBT

Encoding
Total pieces in stem
Total edges in stem
Maximum edges of a piece in stem
Number of curved pieces in stem
Number of pieces with verbal labels in stem
Falsification
Number of pieces falsifiable by gross size, shape and angular disparity
Number of non-falsifiable distractors (NFD)
Number of pieces mismatched by size between NFD and key
Number of pieces mismatched by angle between NFD and key
Expected number of distractors falsified by pieces
Expected number of cycles necessary to falsify NFDs
Confirmation
Number of pieces that must be rotated to match stem to key
Number of pieces displaced between stem and key
Position of key

Participants

The participants for the paper-and-pencil group were 259 undergraduate students from the University of Kansas. The participants for the computer-based group were 91 undergraduate students from the University of Kansas. All

participants were students from the pool of psychology students participating for class credit.

Apparatus and Procedure

Each participant in the first group was given the paper-pencil form of the Revised Minnesota Paper Form Board Test (RMPFBT) Series MA. This test consists of 64 AO items (see Figure 5). Each item is made of a stem and five choices including the correct answer. The stem is a shape that has been cut into varying number of pieces and then separated. The stem item can vary in assembled frame shape (e.g., circle, square, triangle, etc.) and number of pieces (2-5). The pieces can be rotated and/or displaced from their original position. The answer choices consist of one assembled object that is made from the pieces in the stem, whereas the other five alternatives or distractors have the same frame shape but are somehow mismatched by either angular disparity, gross shape size, number of pieces, etc. The participants were given an hour to complete the test.

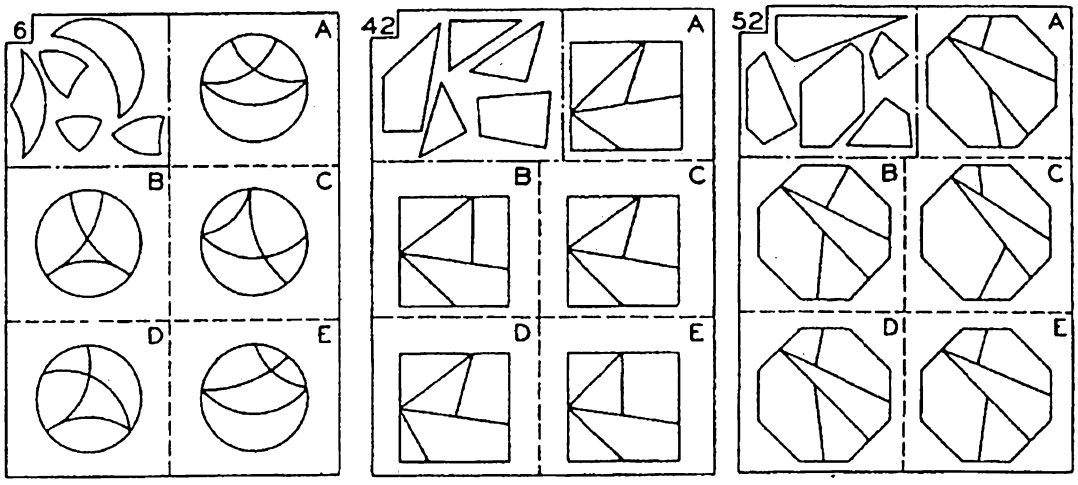

Figure 5. Example RMPFBT items.

The second group of participants were given a computer form of these same items. Each item was presented individually on the computer screen. Participants were not allowed to skip items or return to items they had previously answered. All items were given in the same order as on the paper-and-pencil version of the test. To advance between items, the participant had to use the mouse to choose an answer and then had to click a button labeled "Next". Participants were given the chance the change answers before moving on. The computer program used to administer this test recorded response accuracy and time.

Results

BILOG was used to estimate difficulty for each item based on raw data from the paper-and-pencil test. Log-transformed means were calculated for the response times for all participants where RT was greater than or equal to 3 seconds (any RT

less than 3 seconds implied that the participant did not process the item but rather randomly guessed and moved on). Response times of less than 3 seconds were treated as missing data. Correlations were obtained between the scored variables and item difficulty and mean RT. The program LpcM-WIN 1.0 was used to implement the LLTM to examine the influence of the scored variables on item difficulty. Also, hierarchical regression was used to confirm the influence of the scored variables on item difficulty and RT.

Descriptive Statistics

The first step in this analysis was to score each item on the list of variables mentioned in the previous section. A summary of these scores can be seen in Table 2.

Item difficulty ranged from a minimum score of –3.901 to a maximum score of 2.335, and a mean near the center of -.013 (SD = 1.392). A complete listing of these item difficulties can be found in Appendix A.

Descriptive statistics for item difficulty and response time were a result of collected data. All other variables listed in Table 2 are scored variables. Of the encoding variables, total edges in the stem had the most variability, ranging from 4 to 24 total edges in the stem (N = 64, M = 12.125, SD = 4.088). Ranging from 0 to 5, very few items had pieces with curved edges (N = 64, M = 0.875, SD = 1.579), which is seen because most items containing pieces with curved edges contain at least 2 to 4 pieces with curved edges. For the falsification stage, this table demonstrates that there are very few items with non-falsifiable distractors (NFD). Out of a possible 4 distractors, the mean expected number falsified by any piece was 2.691 (SD = 0.706).

Table 2.

Descriptive Statistics for RMPFBT Items

Variable	N	Min	Max	Mean	SD
Item Difficulty	64	-3.901	2.335	-0.013	1.392
Mean RT	64	7.742	31.607	18.511	5.421
Transformed Mean RT	64	1.920	3.196	2.734	0.266
Total pieces in stem	64	2.000	5.000	3.516	0.816
Total edges in stem	64	4.000	24.000	12.125	4.088
Max edges in stem	64	2.000	7.000	4.407	1.046
Number of pieces with verbal labels	64	0.000	5.000	2.297	1.122
Number of pieces with curved edges	64	0.000	5.000	0.875	1.588
Expected number of distractors falsified	64	0.000	4.000	2.691	0.706
Number of non-falsifiable distractors (NFD)	64	0.000	2.000	0.219	0.453
Number of mismatched pieces between key and NFD	64	0.000	2.000	0.203	0.510
Number of mismatched angles between key and NFD	64	0.000	1.000	0.063	0.244
Expected number of cycles to falsify NFD	64	0.000	3.000	0.445	0.914
Number of displaced pieces between stem and key	64	0.000	3.000	0.797	0.760
Number of rotated pieces between stem and key	64	0.000	5.000	2.266	1.198

For most items, the majority of distractors were grossly falsifiable. There were exactly 13 items with 1 or 2 NFDs. The actual mean for number of NFDs based on

only the 13 items with NFDs is higher than that listed above (N = 13, M = 1.08, SD = 0.277). The expected number of cycles required to falsify a NFD ranged from 1.33 pieces used to falsify to 3 pieces (N = 13, M = 2.192, SD = 0.485). The two confirmation variables, displacement and rotation of pieces, varied from each other--while the number of pieces displaced between the stem and key ranged from 0 to 3 (N = 64, M = 0.797, SD = 0.760), the number of pieces rotated ranged from 0 to all 5 pieces (N = 64, M = 2.266, SD = 1.198). This shows that many more item stems contained rotation than displacement.

The RT distributions for each item were significantly skewed, and thus, the distributions were transformed using a log transformation before calculating the mean RT for each item. Mean response time ranged from 7.472 seconds to 31.607 seconds with a mean of 18.511 seconds (standard deviation=5.421). The original mean RTs and the transformed mean RTs can be seen in Appendix A.

In Figure 6, we can see a highly positive relationship between item difficulty and RT. A significant positive correlation ($r=.733$, $p<.001$) was found between item difficulty and RT.

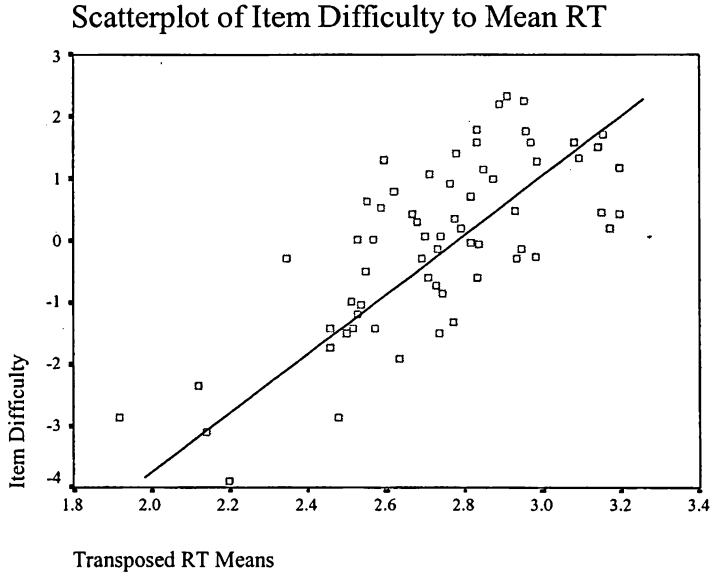

Figure 6. Relation of Item Difficulty with Transformed Mean RT.

Bivariate correlations were run to find which scored variables were significant predictors of item difficulty and mean RT. Table 3 shows these correlations.

Table 3

Correlations of Cognitive Model Variables with Item Statistics

Variable	Item Difficulty	Response Time
Encoding		
Number of pieces in stem	.360**	.658**
Total edges in stem	.255*	.525*
Max edges in stem	.119	.273*
Number of pieces with verbal labels	-.008	-.128
Number of curved pieces	-.173	-.082
Falsification		
Expected number of distractors falsified	-.281*	-.274*
Number of non-falsifiable distractors (NFD)	.277*	.238
Number of mismatched pieces between key and NFD	.182	.247*
Number of mismatched angles between key and NFD	.220	.080
Expected number of cycles to falsify NFD	.350**	.302*
Confirmation		
Number of displaced pieces	.277*	.492**
Number of rotated pieces	.541**	.686**
Key position	-.078	-.155

Note. *$p<.05$ **$p<.01$

Table 3 shows that item difficulty is significantly correlated with the number of pieces in the stem ($r = .360$, $p < .01$), total edges in the stem ($r = .255$, $p < .05$), expected number of pieces falsified ($r = -.281$, $p < .05$), number of non-falsifiable distractors ($r = .277$, $p < .05$), expected number of cycles to falsify a non-falsifiable distractor ($r = .350$, $p < .01$), number of pieces displaced between the stem and key ($r = .277$, $p < .05$), and the number of pieces rotated between the stem and key ($r = .541$,

$p < .01$). Table 3 shows that response time is significantly correlated with the number of pieces in the stem ($r = .658$, $p < .01$), the total number of edges in the stem ($r = .525$, $p < .01$), the maximum number of edges in one piece of the stem ($r = .273$, $p < .05$), the expected number of distractors falsified ($r = -274$, $p < .05$), the number of mismatched pieces between key and non-falsifiable distractor ($r = .247$, $p < .05$), expected number of cycles to falsify a non-falsifiable distractor ($r = .302$, $p < .05$), the number of pieces displaced between the stem and key ($r = .492$, $p < .01$), and the number of pieces rotated between the stem and key ($r = .686$, $p < .01$).

Cognitive Models – Multiple Regression

Hierarchical regression models were run using variables that were not highly correlated with each other within a stage as the predictors of mean RT. The variables for each processing stage were blocked together. In Model 1, the following encoding variables were included: total edges in the stem, maximum edges in one piece and number of pieces with verbal labels. In Model 2, the following falsification variables were added to the encoding variables: expected number of distractors falsified and expected number of cycles to falsify a non-falsifiable distractor. In Model 3, the full model, the following confirmation variables were added: number of displaced pieces between the key and stem and number of rotated pieces between the key and stem.

Table 4

Model Summary for Response Time as the Dependent Variable

					Change Statistics				
Model	R	R^2	Adjusted R^2	SE	R^2 change	F change	df1	df2	Sig.
1	.693[a]	.481	.455	.196	.481	18.525	3	60	.000
2	.712[b]	.507	.465	.195	.026	1.539	2	58	.223
3	.759[c]	.575	.522	.184	.068	4.514	2	56	.015

[a] Predictors: (Constant), Number of pieces with verbal labels, Total pieces in stem, Max edges in stem
[b] Added predictors: Expected number of cycles to falsify NFD, Expected number of distractors falsified
[c] Added predictors: Number of displaced pieces between stem and key, Number of rotated pieces between stem and key

All models were found to be significant—Model 1, $F(3, 60)=18.525$, $p=.000$, Model 2, $F(5, 58)=11.930$, $p=.000$, and Model 3, $F(7, 56)=10.844$, $p=.000$. Table 4 shows that the R^2 changes were significant for the first and third models. This suggests that the falsification variables do not contribute much to the prediction of response time. Table 5 shows both total pieces in the stem and the rotation variable significantly contributed to the prediction of response time, while number of pieces with verbal labels somewhat predicted response time.

Table 5

Coefficients for Modeling Response Time

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	SE	Beta	t	Sig.
1	(Constant)	-2.004	.151		13.268	.000
	Total pieces in stem	.218	.032	.667	6.841	.000
	Max edges in stem	.017	.026	.065	.633	.529
	Number of pieces with verbal labels	-.045	.024	-.188	-1.848	.070
2	(Constant)	2.117	.213		9.931	.000
	Total pieces in stem	.196	.034	.602	5.712	.000
	Max edges in stem	.022	.026	.086	.838	.405
	Number of pieces with verbal labels	-.044	.025	-.186	-1.780	.080
	Expected number of distractors falsified	-.030	.040	-.078	-.747	.458
	Expected number of cycles to falsify NFD	.040	.028	.139	1.433	.157
3	(Constant)	2.231	.206		10.853	.000
	Total pieces in stem	.116	.044	.356	2.643	.011
	Max edges in stem	.012	.025	.048	.480	.633
	Number of pieces with verbal labels	-.033	.024	-.139	-1.387	.171
	Expected number of distractors falsified	-.030	.038	-.079	-.789	.433
	Expected number of cycles to falsify NFD	.021	.027	.072	.760	.450
	Number of displaced pieces between key and stem	.011	.040	.031	.274	.785
	Number of rotated pieces between key and stem	.082	.028	.367	2.923	.005

A hierarchical regression was applied in the same manner for item difficulty. The same variables were added for each model as in the mean RT models.

Table 6

Model Summary for Item Difficulty as the Dependent Variable

					Change Statistics				
Model	R	R^2	Adjusted R^2	SE	R^2 change	F change	df1	df2	Sig.
1	.364[a]	.133	.089	1.328	.133	3.064	3	60	.035
2	.477[b]	.227	.161	1.275	.095	3.549	2	58	.035
3	.592[c]	.350	.269	1.190	.123	5.277	2	56	.008

[a] Predictors: (Constant), Number of pieces with verbal labels, Total pieces in stem, Max edges in stem
[b] Added predictors: Expected number of cycles to falsify NFD, Expected number of distractors falsified
[c] Added predictors: Number of displaced pieces between stem and key, Number of rotated pieces between stem and key

All three models were found to be significant—Model 1, $F(3, 60)=3.064$, $p=.035$; Model 2, $F(5, 58)=3.414$, $p=.009$; and Model 3, $F(7, 56)=4.306$, $p=.001$. The R^2 changes for Models 1, 2 and 3 were significant. Table 6 shows the R-square statistics. Table 7 shows that only total number of pieces in stem and rotation proved to be significant contributors.

Table 7

Coefficients for Modeling Item Difficulty

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	SE	Beta	t	Sig.
1	(Constant)	-2.194	1.021		-2.149	.036
	Total pieces in stem	.612	.215	.359	2.848	.006
	Max edges in stem	.037	.178	.028	.209	.835
	Number of pieces with verbal labels	-.053	.163	-.043	-.324	.747
2	(Constant)	-1.043	1.396		-.747	.458
	Total pieces in stem	.401	.225	.235	1.779	.080
	Max edges in stem	.091	.172	.068	.527	.600
	Number of pieces with verbal labels	-.048	.162	-.039	-.300	.766
	Expected number of distractors falsified	-.302	.259	-.153	-1.166	.248
	Expected number of cycles to falsify NFD	.398	.185	.261	2.155	.035
3	(Constant)	-.318	1.330		-.239	.812
	Total pieces in stem	-.109	.264	-.064	-.385	.702
	Max edges in stem	.035	.164	.027	.216	.830
	Number of pieces with verbal labels	.016	.154	.013	.103	.918
	Expected number of distractors falsified	-.327	.245	-.166	-1.366	.187
	Expected number of cycles to falsify NFD	.264	.178	.174	1.489	.142
	Number of displaced pieces between key and stem	-.066	.258	-.036	-.257	.798
	Number of rotated pieces between key and stem	.585	.181	.504	3.242	.002

Cognitive Models - LLTM

Cognitive model parameter estimates were obtained through the logistic linear latent trait model analysis. The same variables were used for modeling item difficulty here as above. Model 1 consisted of total number of pieces in the stem, maximum number of edges in one piece in the stem, and number of pieces with verbal labels. Model 2 added two variables—expected number of distractors falsified and expected number of cycles to falsify a non-falsifiable distractor. Model 3 consisted of the variables from Model 1 and the added variables in Model 2 as well as the number of pieces displaced between the stem and the key and the number of pieces rotated between the stem and the key.

Table 8

LLTM Model Summary for Item Difficulty as the Dependent Variable

Model	-2lnL	χ^2	Delta df	Delta	fit index Delta $^{1/2}$
Null	15529.72		1		
1	15164.01[a]	365.71	3	.118	.344
2	14854.03[b]	309.98	2	.217	.470
3	14240.95[c]	613.08	2	.414	.640
Saturated	12422.28[d]	1818.67	56		

[a] Predictors: (Constant), Number of pieces with verbal labels, Total pieces in stem, Max edges in stem
[b] Added predictors: Expected number of cycles to falsify NFD, Expected number of distractors falsified
[c] Added predictors: Number of displaced pieces between stem and key, Number of rotated pieces between stem and key
[d] Rasch model

Table 8 shows the fit index for the three models being tested. Delta is defined as the difference between the likelihoods of the model being tested and the Null Model divided by the difference between the likelihoods of the Null Model and

the Saturated Model. This table shows that as the variables are added, the model fit index increases. These fit indices are almost equal to the R^2 from the regression models of item difficulty. This supports the previous data suggesting that the variables are contributing to the prediction of item difficulty.

Table 9

LLTM Coefficients for Modeling Item Difficulty

Model		B	SE	z
1	Total pieces in stem	.445	.025	17.932**
	Max edges in stem	.024	.021	1.157
	Number of pieces with verbal labels	-.007	..018	-.324
2	Total pieces in stem	.285	.027	10.547**
	Max edges in stem	.071	.021	3.326**
	Number of pieces with verbal labels	-.010	.019	.491
	Expected number of distractors falsified	-.271	.031	8.644**
	Expected number of cycles to falsify NFD	.285	.021	13.639**
3	Total pieces in stem	-.264	.041	6.417**
	Max edges in stem	.036	.023	1.591
	Number of pieces with verbal labels	.065	.020	3.167**
	Expected number of distractors falsified	-.336	.033	10.188**
	Expected number of cycles to falsify NFD	.185	.021	8.628**
	Number of displaced pieces between key and stem	-.130	.034	3.823**
	Number of rotated pieces between key and stem	.612	.025	24.102**

Note. *$p<.05$ **$p<.01$

Table 9 shows the contributions of the variables to the model according to the LLTM. According to this table, all predictors contributed significantly to item difficulty with the exception of number of pieces with verbal labels.

Discussion

Through this study, variables were defined and measured that might affect the difficulty and response time of an Assembling Object (AO) item. These variables were not only defined individually, but also defined in terms of the processing stage they influence in the cognitive model. While many of these variables were highly intercorrelated with other variables within the same processing stage, certain variables were found to be highly correlated with item difficulty and response time.

The encoding variables, number of pieces in stem, total edges in stem, and maximum number of edges in one piece in the stem were all significantly positively correlated with either item difficulty or response time. This suggests that as the values of these variables increase, so does item difficulty and the amount of time it takes a subject to solve an item. This supports the postulated encoding stage of the processing model.

The falsification variables, expected number of cycles to falsify a non-falsifiable distractor and number of non-falsifiable distractors were both significantly correlated with item difficulty, suggesting that as the number of non-falsifiable distractors and the number of cycles necessary to falsify those non-falsifiable distractors increase, so did item difficulty. On the other hand, expected number of cycles to falsify a non-falsifiable distractor and number of pieces mismatched by size between the non-falsifiable distractor and the key were both significantly correlated with response time, while the number of non-falsifiable distractors was marginally correlated with response time. Again, this suggests that as these variables increase, so

does response time. Also in falsification, the expected number of distractors falsified was significantly negatively correlated with both response time and item difficulty. This means that as the number of distractors falsified increases, both response time and item difficulty decrease. Taken together, the pattern of results supports the postulated falsification stage.

Finally, the confirmation variables, number of displaced pieces between the stem and the key and the number of pieces rotated between the stem and the key both had very significant positive correlations with item difficulty and response time. Thus, as the number of pieces rotated or displaced between the stem and the key increases, the item difficulty and response time increase as well. Thus, the postulated confirmation stage is supported.

Variables from among those with significant correlations with item difficulty and response time that were not highly intercorrelated were selected to produce the models tested in order to see which of these variables were significant predictors of item difficulty and response time. The encoding variables used for this measurement included number of pieces in stem, maximum number of edges in one piece in the stem and number of pieces with verbal labels. The falsification variables included were the expected number of distractors falsified and the expected number of cycles necessary to falsify a non-falsifiable distractor. The confirmation variables used included number of pieces rotated and number of pieces displaced between the stem and the key. Model 1 consisted only of encoding variables. Model 2 tested the

influence of the encoding variables and the falsification variables. And, Model 3 tested all variables for all three stages.

Through hierarchical regression, it was found that the R^2 for each model was significant. The change between the three models was also significant. This suggests that each set of variables does in fact significantly contribute to the prediction of item difficulty as well as to response time. But, looking at the individual coefficients, we see that only number of pieces in the stem and the number of pieces rotated between the stem and the key seem to contribute significantly to the prediction of item difficulty and response time. Number of pieces with verbal labels marginally influences the prediction of response time as well. Modeling through regression does have its limitations. The standard errors tend to be too large because they depend on the number of items rather than the number of subjects. This results in too few significant predictors.

Through linear logistic latent trait modeling, similar results are found. Again, it was found that the change between models was significant and each model seems to have a high fit index suggesting that each stage of the process contributes significantly to the prediction of item difficulty. Due to the greater accuracy gained when using raw data as well as the scored variable matrix, LLTM found almost all variables to significantly contribute to the prediction of item difficulty. The only variable which did not significantly contribute was number of pieces with verbal labels. This variable also did not individually correlate significantly with item difficulty.

This data suggests that by varying things in the stem such as the number of pieces, number of edges on those pieces and maybe number of pieces with verbal labels, one could influence the difficulty level of the item as well as the amount of time required to solve the item. By varying the number of non-falsifiable distractors, one could greatly affect the item difficulty and response time, because the number of non-falsifiable distractors highly influences the two variables measured under the falsification stage. Remember, though, that in this sample of items, there were only 13 items with non-falsifiable distractors. The small sample of this item type limits its predictive power. Finally, by simply rotating and displacing the pieces in the stem in comparison to those in the key, item difficulty and response time are most easily manipulated. The unequal impact of rotation and displacement may result from the great difference between the means on these two variables. Many more items had pieces rotated than displaced. And, every item that had displacement also had rotation. Although the data here suggests that rotation has a greater influence, if there were more items with displacement to be tested, different results may be found.

Embretson and Gorin (2001) found similar results in their study of AO items. Though their items were similar they were not structured exactly the way the RMPFBT items are, the stem and alternatives were very similar individually. Their results were similar to those found in this study. Most of the same scored variables were significantly correlated with both response time and item difficulty. Embretson and Gorin also found similar results in their regression analysis of the proposed processing model. Though they divided the confirmation stage into two parts, those

related to this study did contribute significantly to the prediction of both item difficulty and response time.

Study 2

Methods

Participants

The participants were 10 students from the University of Kansas. All participants were volunteers.

Apparatus and Procedure

The group of participants was given all 64 of these same RMPFBT items. These participants were first given a computer administered version of 54 of the 64 items. These items were administered in the same order and manner as the previous group. The other ten items were chosen from the total 64 to cover the range of item difficulty, number of pieces, number of non-falsifiable distractors, and stem shape. These items were administered using an eye tracker program.

For the eye tracker data, participants were placed in a chair approximately two feet from the computer screen displaying each item individually. Participants had to wear a visor with a head tracker connected near the left eye. An eye tracker camera was positioned below the computer monitor and was calibrated to focus on the left eye. The eye tracker followed the participants left pupil as they scanned each item. The participant was required to answer the question orally and then click the space bar to proceed to the next item. Responses were recorded by a lab assistant.

Results

Descriptive Statistics

The following table (Table 10) shows the variance on the scored variables of the items used for the eye tracker study. Items were chosen as a representative sample of all 64 items. The parameters for the first sample from Study 1 are listed in parentheses underneath the Study 2 sample parameters for comparison purposes. As Table 10 shows, the range, means and standard deviations for the sample of 10 items are quite representative of the entire item population. The only large discrepancy is found between the sample and population parameters for one variable—expected number of cycles necessary to falsify a non-falsifiable distractor.

Table 10

Descriptive Statistics for 10 Eye Tracker Items

Variable	N	Min	Max	Mean	SD
Total pieces in stem	10	2.000	5.000	3.900	0.994
	(64)	(2.000)	(5.000)	(3.516)	(0.816)
Total edges in stem	10	6.000	24.000	14.500	5.642
	(64)	(4.000)	(24.000)	(12.125)	(4.088)
Max edges in stem	10	3.000	6.000	4.200	1.033
	(64)	(2.000)	(7.000)	(4.407)	(1.046)
Number of pieces with verbal labels	10	0.000	4.000	2.000	1.155
	(64)	(0.000)	(5.000)	(2.297)	(1.122)
Number of pieces with curved edges	10	0.000	5.000	0.900	1.912
	(64)	(0.000)	(5.000)	(0.875)	(1.588)
Expected number of distractors falsified	10	2.000	3.500	2.627	0.505
	(64)	(0.000)	(4.000)	(2.691)	(0.706)
Number of non-falsifiable distractors (NFD)	10	0.000	1.000	0.400	0.516
	(64)	(0.000)	(2.000)	(0.219)	(0.453)
Number of mismatched pieces between key and NFD	10	0.000	1.000	0.400	0.516
	(64)	(0.000)	(2.000)	(0.203)	(0.510)
Number of mismatched angles between key and NFD	10	0.000	0.000	0.000	0.000
	(64)	(0.000)	(1.000)	(0.063)	(0.244)
Expected number of cycles to falsify NFD	10	0.000	3.000	1.050	1.383
	(64)	(0.000)	(3.000)	(0.445)	(0.914)
Number of displaced pieces between stem and key	10	0.000	2.000	0.700	0.675
	(64)	(0.000)	(3.000)	(0.797)	(0.760)
Number of rotated pieces between stem and key	10	1.000	5.000	2.500	1.354
	(64)	(0.000)	(5.000)	(2.266)	(1.198)

By looking at the mean response times and average percent of subjects who answered each item correctly, we can see that the small sample (N = 10) used for the eye tracker study was representative of the full sample used in the other two samples

($N_1 = 259$, $N_2 = 91$). Table 11 shows these statistics. Note that the percent of subjects to answer each item correctly was used in place of item difficulty due to sample size. Since item difficulty means an item is harder to answer correctly, item difficulty was almost perfectly negatively correlated with percent correct for Study 1 ($r = -.945$, $p < .001$).

Table 11

Item difficulty and response time statistics for Study 2 (Study 1 statistics are found in parentheses)

Variable	N	Min	Max	Mean	SD
Mean RT	10	5.800	56.400	25.988	10.736
	(64)	(7.470)	(31.610)	(18.511)	(5.421)
Transformed Mean RT	10	1.730	3.880	3.027	0.424
	(64)	(1.920)	(3.200)	(2.734)	(0.266)
Percent of subjects to answer correctly	10	40.000	100.000	82.500	16.523
	(64)	(36.540)	(98.455)	(74.670)	(17.067)

Individual participant data for all 64 items can be seen in Table 12. When all items are taken into account, we see that percent correct ranges from 64.1% to 98.4% ($N = 10$, $M = 82.500$, $SD = 11.595$) and mean subject response time ranges from 14.609 seconds to 48.141 seconds ($N = 10$, $M = 25.988$, $SD = 9.798$).

Table 12

Individual participant RTs and percent correct for eye tracker study

Group	Participant	Percent correct	Mean RT (N_{items}=64)
High Ability	1	82.813	20.547
	2	98.438	34.641
	3	90.625	23.688
	7	95.313	27.313
	8	85.938	22.844
	10	87.500	25.984
Low Ability	4	65.625	48.141
	9	79.688	27.125
Guessers	5	64.063	14.984
	6	75.000	14.609

Qualitative Analyses

Eye scan paths and fixation points were analyzed qualitatively with respect to the proposed processing model. First, participants were divided into three groups—Group 1 consisted of high ability participants; Group 2 consisted of low ability participants; and Group 3 consisted of guessers. High ability participants had a high overall percent correct scores (N = 6, M = 90.105, SD = 5.902) and an average to above average mean response time (N = 6, M = 25.836, SD = 4.925). Low ability participants had a low overall percent correct scores (N = 2, M = 72.656, SD = 9.944) and an average to above average mean response time (N = 2, M = 37.633, SD = 14.861). Guessers tended to have a low overall percent score (N = 2, M = 69.532, SD = 7.734) as well as a below average mean response time (N = 2, M = 14.797, SD = 0.265).

After looking over three dimensional graphs of the amount of time spent in any particular look zone, three participants were selected to demonstrate the above mentioned group tendencies. Figures 7 through 9 show in bar graph form the amount of time proportional to other lookzones that each subject spent on a particular look zone on item 55 (see Figure 6).

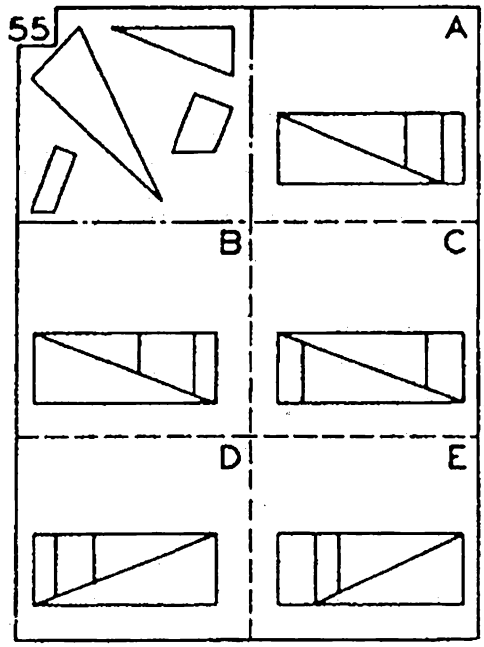

Figure 7. Item 55 of the RMPFBT.

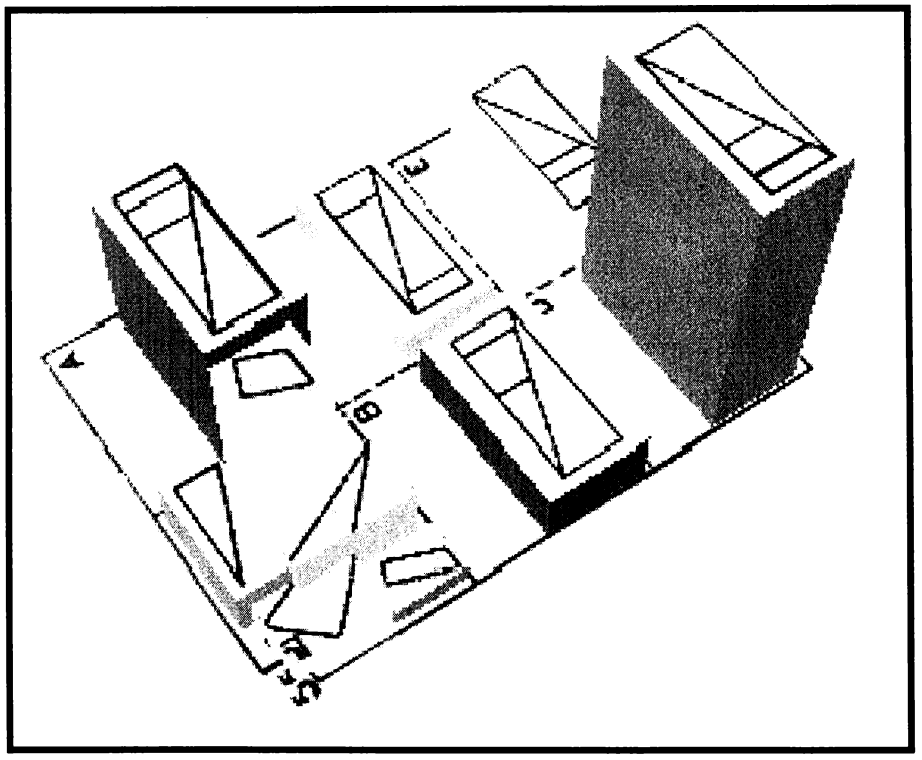

Figure 8. Three-dimensional portrayal of time spent on each area of item 55 for a high ability participant: 2 (score = 98.4, mean RT = 34.641).

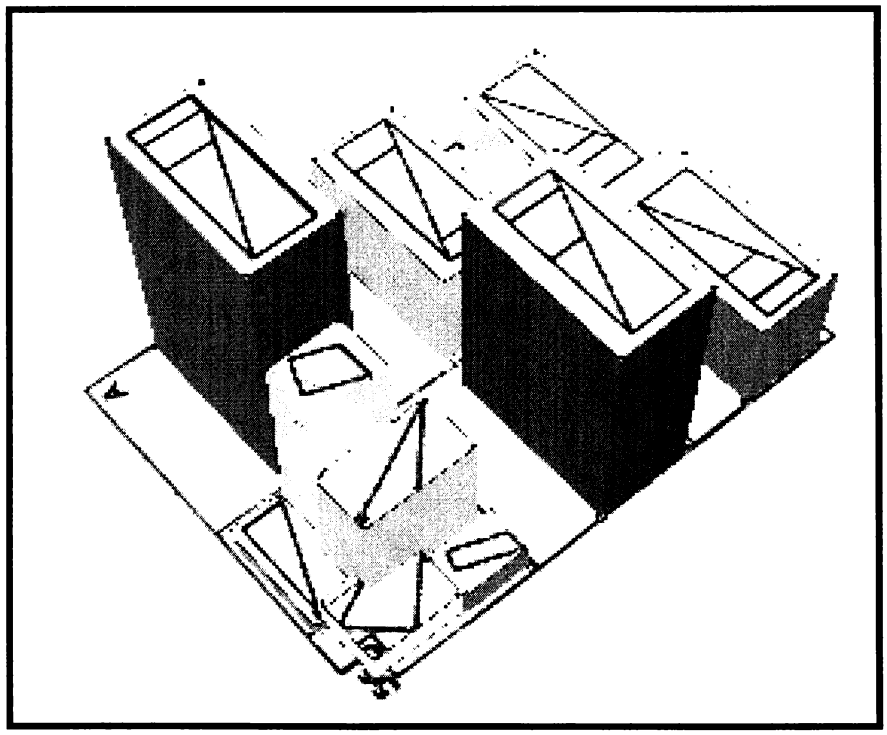

Figure 9. Three-dimensional portrayal of time spent on each area of item 55 for a low ability participant: 4 (score = 65.6, mean RT = 48.141).

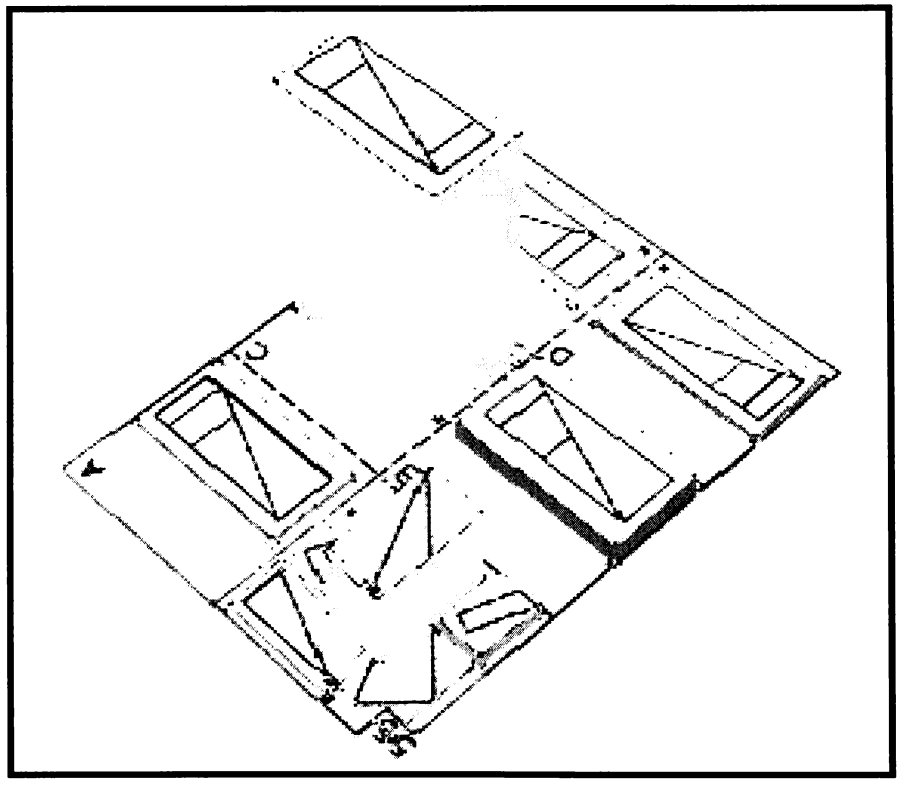

Figure 10. Three-dimensional portrayal of time spent on each area of item 55 for a guesser: 6 (score = 75.0, mean RT = 14.609).

The correct answer for item 55 is choice D as can be seen in Figure 6. Subject 2, the high ability subject, gave the correct answer. Subject 2 also spent a proportional amount of time on choice A. The strategy used by Subject 2 seems to be to choose one piece from the stem and falsify it with each alternative, as can be seen by the one piece that has a much higher bar than the other four pieces. Subject 4, the low ability subject, did not answer this question correctly. He answered choice C. Subject 4 chose the same piece to falsify but spent a considerable amount of time on all alternatives before making a decision, which possibly could have been a guess as well. Subject 6, the guesser, on the other hand spent no time on the pieces and very little time on any alternatives besides choice C, which was incorrect. These patterns were seen throughout all 10 items over all 10 subjects.

Discussion

The eye scan paths seem to support our proposed three stage model. It seems that rather than encoding all stem pieces before beginning the falsification stage, some participants might pick one particular piece and begin the falsification stage after spending some time encoding that particular piece and only briefly scanning the other pieces in the stem. Because there are so few items with non-falsifiable distractors, if a participant picks the best piece on his first try, he can very simply come to find the key among the distractors. With the item that did contain a non-falsifiable distractor among the items in the eye tracker portion of this study, most participants in the high ability and low ability group spent some time encoding one piece and then falsifying all falsifiable distractors first. Then, they began encoding

more pieces and doing more comparisons between the other pieces and the two alternatives left to decide between.

Conclusion

Study 1 demonstrates that variables can be defined that affect item difficulty and response time for these item types. This enables test developers to design varying items based on these attributes effectively affecting the difficulty level of the test overall. Study 1 also demonstrated that the most important variables affecting item difficulty and response time for items of this type are the number of pieces in the stem, the number of non-falsifiable distractors, and the amount of mental rotation required to confirm the key.

The models from Study 1 supported the proposed cognitive processing model by dividing the variables into the three processing stages: encoding, falsification and confirmation. Each block of variables contributed significantly to the prediction of item difficulty and response time. This data supports theories as to which variables affect which processing stages for solving assembling object items.

Qualitative analyses from Study 2 seemed to suggest that participants pick one particular piece and use it to falsify alternatives. This data tends to indicate that the three stage model does work, but that each stage is not necessarily fulfilled completely before moving on to the next stage due to the amount of falsification available in this particular test form of this item type.

Further research on these item types could be conducted to more accurately delineate the amount of time taken in each processing stage to help decide the exact

strategy used by different ability level participants in solving this particular item type. By breaking down response time, researchers could better perfect computer generation of these assembling object items by being able to vary attributes affecting item difficulty.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Banich, M. T. (1997). *Neuropsychology: The neural bases of mental function.* New York: Houghton Mifflin Company.

Bertera, J. H. & Rayner, K. (2000). Eye movements and the span of the effective stimulus in visual search. *Perception & Psychophysics, 62(3),* 576-585.

Binet, A., & Simon, T. (1916). *The development of intelligence in children* (E. S. Kite, Trans.). Baltimore: Williams & Wilkens.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cooper, L. A., & Shephard, R. N. (1973). Choronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing.* New York: Academic Press.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3 (3)*, 380-396.

Embretson, S. E. (1999). Cognitive Psychology Applied to Testing. In F. T. Durso, R. S. Nickerson, R. W. Schvaneveldt, S. T. Dumais, D. S. Lindsay & M. T. H. Chi (Eds.), *Handbook of Applied Cognition.* New York: John Wiley & Sons Ltd.

Embretson, S. E. & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38 (4)*, 343-368.

Embreston, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *MMSS: Fundamentals of item response theory*. Newbury Park, CA: SAGE Publications.

Hoffman, J. E. (1998). Visual attention and eye movements. In H. Pashler (Ed.), *Attention*. Boulder, CO: NetLibrary.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57(6),* 787-795.

Humphreys, L. G., Lubinski, D., & Yao, G. (1993). Utility of predicting group membership and the role of spatial visualization in becoming an engineer, physical scientist, or artist. *Journal of Applied Psychology, 78(2),* 250-261.

Hunt, E. (1982). Towards new ways of assessing intelligence. *Intelligence, 6*, 231-240.

Just, M. A. & Carpenter, P. A. (1985). Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability. *Psychological Review, 92(2)*, 137-171.

Kosslyn, S. M. (1980). *Image and mind.* Cambridge, MA: Harvard University Press.

Lansman, M., Donaldson, G., Hunt, E., & Yantis, S. (1982). Ability factors and cogntive processes. *Intelligence, 6*, 347-386.

Likert, R., & Quasha, W. H. (1970). *Manual for the Revised Minnesota Paper Form Board Test*. New York: The Psychological Corporation.

Lohman, D. F. (1979). *Spatial ability: A review and reanalysis of the correlational literatures* (Tech. Rep. No. 8). Standford, CA: Stanford University, Aptitude Research Project, School of Education. (NTIS No. AD-A075 973).

Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, 4*. Hillsdale, NJ: Erbaum.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.). *Handbook of Intelligence*. New York: Cambridge University Press.

Luce, R. D. (1986). *Response Times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

McGee, M. (1979). *Human spatial abilities: Sources of sex differences*. New York: Praeger.

Mumaw, R. J., Pellegrino, J. W., Kail, Jr., R. V., & Carter, P. (1984). Different slopes for different folks: Process analysis of spatial aptitude. *Memory & Cognition, 12 (5)*, 515-521.

Noton, D., & Stark, L. (1971). Eye movements and visual perception. *Scientific American, 224*, 34-43.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. Cambridge, Massachusetts: The MIT Press.

Pellegrino, J. W., Mumaw, R. J., & Shute, V. J. (1985). Analyses of spatial aptitude and expertise. In S. Embretson (Ed.), *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Salvucci, D. D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. Paper presented at Eye Tracking Research & Applications Symposium 2000. Palm Beach Gardens, FL.

Shephard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Sternberg, R. J. (Ed.) (1994). *Encyclopedia of Human Intelligence, Vol. 2.* New York: MacMillan Publishing Company.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review, 38*, 406-427.

United States Employment Services. (1957). *Estimates of worker trait requirements for 4,000 jobs.* Washington, DC: Government Printing Office.

West, T. G. (1991). *In the mind's eye.* Buffalo, NY: Prometheus Books.

Yarbus, A. L. (1967). *Eye movements and vision.* New York: Plenum.

Appendix A

Item Difficulty and Mean RT for all items

Item ID	Item Difficulty	Mean RT	Transformed Mean RT
1	-2.856	7.47	1.92
2	-3.091	9.62	2.14
3	-1.439	14.76	2.57
4	-1.731	12.58	2.46
5	-1.314	18.68	2.77
6	-1.506	17.33	2.73
7	-2.857	12.86	2.48
8	-1.906	15.55	2.63
9	-2.354	9.06	2.12
10	-3.901	9.91	2.20
11	-1.042	14.34	2.53
12	-1.199	13.56	2.53
13	-0.727	17.37	2.73
14	-1.439	13.81	2.52
15	-0.609	16.69	2.70
16	-0.025	19.05	2.82
17	-0.854	18.56	2.74
18	-0.609	21.38	2.83
19	0.190	28.43	3.17
20	0.003	14.67	2.57
21	-0.295	22.46	2.94
22	-0.264	21.97	2.99
23	-0.296	16.00	2.69
24	-1.439	12.79	2.46
25	-1.506	14.05	2.50
26	0.437	30.81	3.20
27	-0.993	13.76	2.51
28	0.621	14.55	2.55
29	-0.141	18.60	2.73
30	0.003	13.93	2.53
31	-0.054	19.39	2.84
32	-0.141	22.45	2.95
33	-0.296	11.23	2.35
34	0.058	17.68	2.74

Appendix A

Item Difficulty and Mean RT for all items

Item ID	Item Difficulty	Mean RT	Transformed Mean RT
35	0.190	18.15	2.79
36	1.515	27.54	3.14
37	0.921	17.96	2.76
38	0.795	15.47	2.62
39	0.484	21.35	2.93
40	0.460	30.71	3.15
41	0.530	15.26	2.59
42	1.322	27.32	3.09
43	0.341	18.39	2.78
44	0.708	18.60	2.82
45	0.292	16.44	2.68
46	0.437	16.15	2.67
47	0.058	16.43	2.70
48	1.183	29.34	3.19
49	1.784	19.26	2.83
50	1.283	23.05	2.99
51	-0.498	14.18	2.55
52	1.726	31.61	3.16
53	1.003	19.90	2.87
54	1.144	21.43	2.85
55	1.765	22.05	2.96
56	1.573	18.98	2.83
57	1.064	17.87	2.71
58	1.592	25.72	3.08
59	1.303	15.60	2.60
60	1.592	24.23	2.97
61	2.253	22.94	2.95
62	1.400	19.94	2.78
63	2.213	21.41	2.89
64	2.335	22.11	2.91